Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

IS THREE BETTER THAN ONE? SIMULATING THE EFFECT OF REVIEWER SELECTION AND BEHAVIOR ON THE QUALITY AND EFFICIENCY OF PEER REVIEW

Federico Bianchi

Department of Economics and Management
University of Brescia
Via San Faustino 74/B
25122 Brescia, ITALY

Flaminio Squazzoni

Department of Economics and Management
University of Brescia
Via San Faustino 74/B
25122 Brescia, ITALY

ABSTRACT

This paper looks at the effect of multiple reviewers and their behavior on the quality and efficiency of peer review. By extending a previous model, we tested various reviewer behavior, fair, random and strategic, and examined the impact of selecting multiple reviewers for the same author submission. We found that, when reviewer reliability is random or reviewers behave strategically, involving more than one reviewer per submission reduces evaluation bias. However, if scientists review scrupulously, multiple reviewers require an abnormal resource drain at the system level from research activities towards reviewing. This implies that reviewer selection mechanisms that protect the quality of the process against reviewer misbehavior might be economically unsustainable.

1 INTRODUCTION

Peer review is an essential mechanism to ensure the quality of scientific publications. It also directly or indirectly contributes to the regulation of resource allocation in science as it determines who is published in top journals and who is not. Peer review is largely influenced by scientific motivation and behavior as well as by conventions and rules that regulate such a complex, distance and anonymous collaboration between scientists (e.g., Leek, Taub, and Pineda 2011; Lee et al. 2013; Squazzoni, Bravo, and Takács 2013). Unfortunately, the scarcity of data has impeded a systematic analysis of peer review at journals and funding agencies, leading many observers to conclude that this mechanism lacks any empirical or experimental support (e.g., Smith 2006; Alberts, Hanson, and Kelner 2008; Bornmann 2011; Couzin-Frankel 2013). To fill this gap, some scholars have recently started to use computer simulation to look at certain peer review processes to estimate implications of scientist misbehavior for the quality of this process (e.g., Squazzoni and Takács 2011; Allesina 2012; Payette 2012). For instance, Thurner and Hanel (2011) studied the effect of peer review bias on publication quality by modeling author and reviewer interaction. They tested different possible reviewer behavior, including rational cheating, that is when certain reviewers could try to disqualify the scientific work of colleagues who are more productive. Their results showed that even a small fraction of strategic, unfair reviewers is sufficient to reduce the quality of published work, making even random chance selection better (see also Roebber and Schultz 2011).

Squazzoni and Gandelli (2012) corroborated this finding by looking at the strategic behavior of reviewers in situations where evaluation standards are weak. The lack of standards in the community, typical of social sciences, can even exacerbate the impact of unfair behavior, limiting cumulative reputational advantages by more productive scientists. In a subsequent work, in which they looked at possible reciprocity strategies between authors and reviewers (i.e., if previously published, scientists

behave fairly when cast as reviewers, otherwise they do not), they found that reciprocity can be detrimental as it tends to perpetuate bias.

In these cases, the evaluation bias of peer review is higher than by mere chance. They found that reciprocity motivation is positive only if reviewers do not consider the fate of previous submission but they reciprocate the pertinence of reviewers' judgment they had been exposed to when previously authors (Squazzoni and Gandelli 2013). More recently, Grimaldo and Paolucci (2014) suggested that assigning more than one reviewer could reduce the impact of strategic behavior as does the implementation of different mechanisms of evaluation scoring (e.g., from recommendations to ratings) (e.g., Paolucci and Grimaldo 2014).

Our aim here is to contribute to the scope of this field. On the one hand, we have extended Squazzoni and Gandelli's model (2012; 2013) to look at the impact of multiple reviewers. On the other hand, unlike previous studies, such as Thurner and Hanel (2011) and Grimaldo and Paolucci (2014), we have also looked at resource allocation concerns in order to consider not only the quality but also the economic sustainability of peer review. Indeed, these aspects are increasingly considered when evaluating the pros and cons of different mechanisms, including alternatives to peer review (e.g., Birukou et al. 2011).

The rest of the paper is as follows. Sect. 2 presents the model, while Sect. 3 illustrates varying simulation scenarios that we have built to test manipulations of certain parameters. Sect. 4 presents the results with particular interest on measuring the bias and efficiency of different scenarios. Sect. 5 briefly summarizes the main contribution of this paper and discusses its limitations.

2 THE AGENT-BASED MODEL

Following Squazzoni & Gandelli (2012; 2013), we have assumed a population of N scientists (N = 240) who were called to submit (*authors*) or review (*reviewers*) articles. At each simulation tick, the task for each *author* was to submit an article for publication, while *reviewers* were assigned to one author each in order to evaluate the quality of their submissions. The roles were randomly alternated in each simulation tick.

We assumed that resources were needed both to submit and review an article. Each scientist had a variable amount of resources $R_a \in N$, which was initially set at 0. In each simulation, scientists were endowed with a fixed amount of resources, equal for all (common access to research infrastructure and internal funds, availability of graduate students). Then, they accumulated resources according to their publication score. The more scientists published, the more resources they had. Resources reflected the scientist's academic status and reputation in the community.

We assumed that the expected quality of author submissions (μ_{ar}) depended on a scientist's resources, according to the following equation:

$$\mu_{ar} = \frac{v \cdot R_a}{v \cdot R_a + 1}$$

where v was the velocity at which the quality of a submission varied according to the increase of the author's resources (v = 0.1). Then, the actual quality of submissions by authors proportionally varied from the scientists' expected quality, by following a normal distribution $N(R_a, \sigma)$ (Squazzoni and Gandelli 2013).

The chance of being published was determined by the average evaluation score assigned by reviewers (see below). Depending on the reviewers' opinion, only the best 30 submissions were published in each tick of the simulation. This was to mimic a selective environment with a set of fixed publication opportunities over time, e.g., a restricted and stable number of top journals. If not published, authors lost all resources invested in the submission process.

We assumed that successful publication multiplied author resources by a value M, which varied from 1, in the case of highly productive scientists, to 1.5, in the case of less productive ones, as they gained

more from publishing. More specifically, at the end of each tick, we graded all n published authors per resources according to the A_i sequence, where A_0 was the published author with the least resources and $A_{(n-1)}$ the published authors with more resources. We defined g as the highest multiplier value possible (g = 1.5). The multiplier M for the i-agent was calculated as follows:

$$M_{(A_i)} = g - (\frac{i}{n-1}) * (g-1) \; = \; \frac{g(n-i-1) \; + i}{n-1}$$

We modeled reviewing as a resource-demanding activity. When selected as reviewers, scientists needed to invest a given amount of resources (see below) for reviewing, while simultaneously losing the opportunity to publish. Moreover, the amount of resources spent for reviewing depended on the distance between the quality of the submission to be reviewed and the expected quality of the reviewer as author. The total expense S for each reviewer was calculated as follows:

$$S = \frac{1}{2} R_r [1 + (Q_a - \mu_r)]$$

where R_r was a reviewer's resources, Q_a was the real quality of the submission to be reviewed, and μ_r was the reviewer's expected quality. Furthermore, we assumed that reviewing expenses grew linearly with the quality of an author's submissions. Reviewers spent less when matched with lower quality author submissions, more when matched with higher quality submissions. In addition, reviewing expenses increased proportionally on the scientist's productivity. This meant that top-scientists wasted less time for reviewing in general, as they have more experience and ability to evaluate good science than average scientists have. However, they will lose more resources than average scientists because their time is more costly.

Table 1 shows the simulation parameters. The scientists' resources were set at the beginning of the simulation at 0 for all. At the first tick, 50% of agents were published randomly. Subsequently, each scientist had a fixed productivity gain each tick. Those published had the value of their publication multiplied by the parameter M [1, 1.5].

Table 1: The Simulation Parameters. The values of "unreliability probability" parameters were manipulated only in the "*fair*" and "*random*" scenarios, since in the "strategic" scenario, the reviewers' unreliability endogenously depends on interaction among agents.

Parameters	**Value**
Initial scientist's resources	0
Fixed productivity gain	1
Number of accepted publications	30
Highest publication productivity multiplier	1.5
Unreliability probability	[0, 0.25, 0.33, 0.5]
Number of reviewers per author	[1, 2, 3]
Evaluation bias by default	0.1
Author investment for publication	1
Reviewing expenses of unreliable reviewers	0.5
Underrating by unreliable reviewers	0.1
Overrating by unreliable reviewers	1.9
Velocity of best quality approximation	0.1

3 SIMULATION SCENARIOS

We tested the impact of scientists' behavior on the quality and efficiency of peer review by varying the simulation scenarios. Following Squazzoni and Gandelli (2013), for quality, we meant the capability of peer review to ensure that only the best submissions were eventually published. This was calculated by measuring the difference between the optimal situation, i.e., the ranking of the expected quality of each author in each simulation tick, with the actual situation, that is the ranking given by reviewers ("*evaluation bias*"). We calculated a percentage of error among the list of 30 published submissions in each simulation tick, i.e., the percentage of accepted submissions that ideally should have been rejected. By efficiency, we meant the capability of peer review to minimize resources wasted by reviewers and the best authors who were not published. This was measured by calculating the reviewing expenses as the percentage of resources spent by scientists for reviewing compared with the resources invested by authors to submit ("*reviewing expenses*"). We also measured the percentage of resources wasted by (unpublished) best authors compared with the optimal situation and an inequality index for resource distribution.

In the first scenario (called "*fair*"), we assumed that when selected as reviewers, scientists behaved fairly by investing resources to provide a pertinent opinion on the quality of author submissions. For fairness, we meant the intention of a reviewer to provide a consistent and unequivocal opinion that truly reflected the quality of an author's submission. In this case, we assumed a normal distribution of the reviewers' expected capability of correctly evaluating a submission, which depended on their productivity, as well as a narrow standard deviation of their evaluation score from the real quality of the submission ($\sigma = R_a/100$). This meant that the evaluation scores by fair reviewers were likely to approximate the real value of author submissions. However, we also assumed that reviewers could make mistakes that increased proportionally to the difference between reviewers' expected quality and author submission quality, also in the "*fairness*" scenario. This scenario was used as a baseline to compare the following ones.

In the second scenario (called "*random*"), we assumed that, when selected as reviewers, scientists had a fixed probability of behaving unreliably, which remained constant over time and was not influenced by past experience. In this case, reviewers could randomly fall into type I and type II errors: recommending to publish submissions of low quality or recommending not to publish submissions that should be published. If unreliable, reviewers spent less resources than reliable reviewers (i.e., - 50% of what was ideally needed if they were fair), and under- or over-estimated author submissions. In the third scenario (called "*strategic*"), we assumed that scientists, when selected as reviewers, were influenced by the outcome of their own previous submission, independent of the identity of the reviewers they had been assigned to. In cases in which their past submissions had been accepted, they reciprocated by providing reliable opinions when selected as reviewers. In cases of past rejection, they reciprocated negatively by providing biased opinion on the quality of author submissions. In this case, we assumed that reviewers invested more resources for reviewing as they aimed at increasing the chance that authors were rejected mirroring what had occurred to them previously.

Finally, we manipulated the number of reviewers assigned to each author submission, from 1 to 3, for each scenario. In cases of multiple reviewers, the evaluation of the submission was the average value of the *n* scores provided by reviewers. Furthermore, we also tested different values of the probability of unreliability, 0 being for the *fair* scenario mentioned above.

4 RESULTS

Tab. 2 shows the impact of multiple reviewers in the *fair* and *random* scenario, in which we manipulated reviewer unreliability. While three reviewers had a negative role when reviewers were all fair, there was an inherent bias of judgment due to random noise. Multiple reviewers had a positive effect in reducing bias when reviewers were randomly reliable (=0.50). In this case, the number of errors due to unreliability more than halved when three reviewers were involved instead of one. In cases of strategic behavior by reviewers, the level of bias with one reviewer was 15% higher than when reviewers behaved randomly and 38% higher when reviewers were fair. Selecting more than one reviewer was beneficial to reduce distortion induced by strategic reciprocity, although in general this did not rule out the negative effect of this strategy.

Table 2: The impact of unreliability by reviewers and multiple reviewers on the evaluation bias of peer review with multiple reviewers (values in percentage, averaged over 3,000 simulation runs, $t = 200$).

Degree of unreliability of reviewers	*Number of reviewers*		
	1	*2*	*3*
0.00 (*fair scenario*)	5.59	9.87	13.41
0.25 (*random scenario*)	15.26	12.97	14.86
0.33 (*random scenario*)	20.95	12.78	13.80
0.50 (*random scenario*)	28.97	15.92	12.92
Strategic scenario	43.32	35.20	25.74

Figure 1 shows the effect of multiple reviewers on evaluation bias in the three scenarios, i.e., with fair (random unreliability = 0), random (random reliability = 0.5), and strategic reviewers. On the one hand, strategic behavior generally resulted in higher bias than random behavior, so confirming previous simulation findings (e.g., Thurner and Hanel 2011; Squazzoni and Gandelli 2013). On the other hand, bias was significantly reduced when three reviewers were involved, thus qualitatively confirming Grimaldo and Paolucci's findings (2014).

Obviously, a significant part of the bias generation mechanism was due to the extent to which reviewers under- or over-rated the quality of author submissions. Not only could reviewers be unreliable,

their opinion could be biased with a different order of magnitude. In order to look at this, we tested different distance values of the under/overrating parameter from the true value of the quality of a submission. We found that the impact of reviewer bias was higher when the excursion of this parameter was >70% (i.e., + or – 70% of the actual quality of the submission), while the impact of the order of magnitude of this bias was lower when multiple reviewers were involved in the evaluation of a submission.

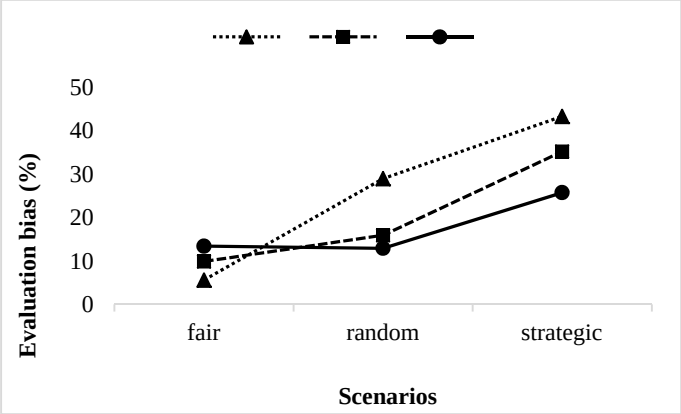

Figure 1: *Evaluation bias* with different reviewer behavior (% values, averaged over 3,000 simulation runs, t = 200).

Table 3 shows the impact of unreliability and multiple reviewers on reviewing expenses. As expected, assigning reviews to multiple reviewers required more resources invested by scientists on reviewing, mostly growing proportionally when reviewers were added. This effort was abnormal in all scenarios but especially when reviewers were all or mostly fair and when they behaved strategically. When reviewers were all fair and evaluating submissions involved three reviewers, it was as if the system were demanding 1.5 resource unit for reviewing for each 1.0 resource authors invested in research (i.e., writing and submitting papers). This is a dramatically unsustainable allocation where system resources are invested more in evaluating than in researching activity. The same happened in the "strategic" scenario, in which reviewing expenses increased from 33.86 to 114 when passing from one to three reviewers. In each combination, random behavior by reviewers was economically more beneficial as it permitted at least half of the scientists cast as reviewers to allocate their resources more in submitting than in evaluating (see Figure 2).

Table 3: The impact of unreliability of reviewers and multiple reviewers on reviewing expenses (values in percentage, averaged over 3,000 simulation runs, t = 200).

Degree of unreliability of reviewers	*Number of reviewers*		
	1	*2*	*3*
0.00 (*fair scenario*)	36.41	93.16	144.54
0.25 (*random scenario*)	25.96	57.81	102.04
0.33 (*random scenario*)	30.13	53.28	93.19
0.50 (*random scenario*)	29.48	51.11	82.92

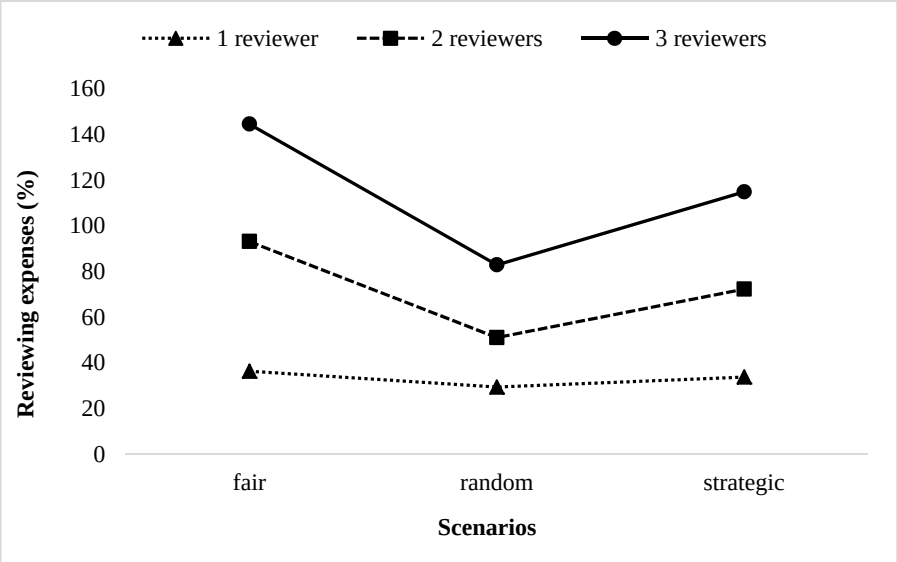

Figure 2: *Reviewing expenses* in the three scenarios (% values, averaged over 3,000 simulation runs, t = 200).

5 CONCLUSIONS

Our findings corroborate previous studies on the effect of reviewer behavior on the quality of peer review (e.g., Thurner and Hanel 2011; Squazzoni and Gandelli 2012, 2013; Grimaldo and Paolucci 2014). These studies indicate that even a minimal difference in reviewer behavior might have significant implications for the quality of publications. The positive effect of multiple reviewers would confirm that the "rule of thumb" used by journal editors to ask opinions from different reviewers is appropriate. On the other hand, while looking at resource allocation, if there are reviewers that are not inspired by fairness and commitment, the quality of peer review comes at a serious cost, i.e., a resource drain from researching to reviewing, which could even reach abnormal levels. This calls for a serious reconsideration of the sustainability of peer review in a phase in which digital publications have exploded and scientists are asked more and more to review not only for journals, but also for book series, funding agencies, and university/department National and local evaluations.

Unfortunately, the lack of data on peer review at journals and funding agencies makes it difficult to test these findings experimentally, e.g., estimating the type and distribution of reviewer behaviors. Given the importance of peer review for science at all levels, e.g., funds, reputation, and innovation, significant improvements will be achieved only when simulation analysis are informed by, calibrated with, and tested against empirical data.

ACKNOWLEDGMENTS

The authors would like to acknowledge support by the TD1306 COST Action on "New Frontiers of Peer Review" and help by Claudio Gandelli on the set up of the model. Usual disclaimers apply.

REFERENCES

Alberts, B., B. Hanson, and K. L. Kelner. 2008. "Reviewing Peer Review." *Science* 321:15.

Allesina, S. 2012. "Modeling Peer Review: An Agent-Based Approach," *Ideas in Ecology and Evolution* 5(2): 27–35.

Birukou A., J. R. Wakeling, C. Bartolini, F. Casati, M. Marchese, K. Mirylenka, N. Osman, A. Ragone, C. Sierra, and A. Wassef. 2011. "Alternatives to Peer Review: Novel Approaches for Research Evaluation," *Frontiers in Computational Neuroscience*.

Bornmann, L. 2011. "Scientific Peer Review." *Annual Review of Information Science and Technology*, 45:199–245.

Couzin-Frankel, J. 2013. "Secretive and Subjective, Peer Review Proves Resistant to Study." *Science* 341: 1331.

Crocker, J. and M. L. Cooper. 2011. "Addressing Scientific Fraud." *Science* 334(6060):1182.

Gewin, V. 2012. "Uncovering Misconduct." *Nature* 485:137–139.

Grimaldo F. and M. Paolucci. 2014. "A Simulation of Disagreement for Control of Rational Cheating in Peer Review," *Advances in Complex Systems*, 99: 663–688.

Lee, C. J., C. R. Sugimoto, G. Zhang, and B. Cronin. 2013. "Bias in Peer Review," *Journal of the American Society for Information Science and Technology* 64: 2–17.

Leek, J. T., M. A. Taub, and F. J. Pineda. 2011. "Cooperation between Referees and Authors Increases Peer Review Accuracy," *PLoS ONE* 6(11): e26897.

Paolucci M. and F. Grimaldo. 2014. "Mechanism Change in a Simulation of Peer Review: From Junk Support to Elitism," *Scientometrics*, 99: 663–688.

Payette, N. 2012. "Agent-Based Models of Science." In Scharnhorst, A., K. Börner and P. Van den Besselaar (Eds.), *Models of Science Dynamics: Encounters between Complexity Theory and Information Sciences*. Springer Verlag, Heidelberg.

Roebber, P. J. and D. M. Schultz. 2011. "Peer Review, Program Officers and Science Funding." *PLoS ONE* 6(4):e18680: http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0018680.

Smith. R. 2006. "Peer Review. A Flawed Process at the Heart of Science and Journals." *Journal of the Royal Society of Medicine* 99: 759–760.

Squazzoni, F., G. Bravo and K. Takács, K. 2013. "Does Incentive Provision Increase the Quality of Peer Review? An Experimental Study." *Research Policy* 42(1): 287–294.

Squazzoni, F. and C. Gandelli. 2012. "Saint Matthews Strikes Again. An Agent-Based Model of Peer Review and the Scientific Community Structures." *Journal of Informetrics* 6:265–275.

Squazzoni, F. and C. Gandelli. 2013. "Opening the black box of peer review. An agent-based model of scientist behaviour." *Journal of Artificial Societies and Social Simulation* 16(2) 3: http://jasss.soc.surrey.ac.uk/16/2/3.html

Squazzoni, F. and K. Takács. 2011. "Social Simulation that 'Peers into Peer Review'." *Journal of Artificial Societies and Social Simulation* 14(4) 3: http://jasss.soc.surrey.ac.uk/14/4/3.html.

Thurner, S. and R. Hanel. 2011. "Peer Review in a World with Rational Scientists: Toward Selection of the Average." *The European Physical Journal B* 84:707–711.

AUTHOR BIOGRAPHIES

FEDERICO BIANCHI is Ph.D. candidate in Economic Sociology and Labor Studies at the University of Milan and Brescia. His field of research is social exchange, social networks, and cooperation. He is a

member of TD1306 COST Action on “New Frontiers of Peer Review” (www.peere.org). His email address is Federico.Bianchi1@unimi.it.

FLAMINIO SQUAZZONI is Associate Professor of Economic Sociology at the University of Brescia, where he leads the GECS-Research Group on Experimental and Computational Sociology (www.eco.unibs.it/gecs). He is editor of the *Journal of Artificial Societies and Social Simulation* and president of ESSA-The European Social Simulation Association. He is Chair of the TD1306 COST Action on “New Frontiers of Peer Review” (www.peere.org). His email address is flaminio.squazzoni@unibs.it.